

School of Computer Science

Machine Learning in Fulfilment of

SPEC9270

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: Sarah Jane Delany

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹

**Date: 2021/03/21**

Acknowledgments

*Collection of this dataset was made at the Mind Research Network under an NIH NIGMS Centers of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 to Vince Calhoun (PI).*

Definition of problem – Kaggle Contest Entered

The Kaggle contest entered is the **MLSP 2014 Schizophrenia Classification Challenge** (Network, 2014), although the competition is over it is still possible to submit work and receive an evaluation. My reasoning for picking this competition is that I wanted to study psychology and go into research particularly in dissociative disorders and schizophrenia, however due to personal circumstances at the time I went with engineering. Data provided are two sets of multimodal features derived from MRI images. Two modalities are used to obtain said features, functional and structural MRI.

Today, I’m still interested in a PhD and have gained a different perspective I otherwise wouldn’t have had, one that sees that there are a lot of holes and short comings in traditional diagnosis of mental disorders such as D.I.D and schizophrenia. All of which hopefully can be dealt with machine learning (ML).

Some of these short comings are that there is no process to date that properly diagnoses D.I.D, despite it being acknowledged as a mental illness in the 1950s, research between then and now has been done but was later found to be fraudulent or difficult to reproduce. (David Blihar, 2020)

While remaining brief and acknowledging other issues that stem from the use of machine learning (ML), MRI and fMRI images can be fed to ML algorithms in order to be used as a tool that can help clinicians avoid misdiagnosis. Young girls tend to not be diagnosed with ADHD because it manifests differently (CONNOLLY, 2020), having a better system for such differences in mental health would help avoid undiagnosed patients in situations such as school performance.

Currently the best/popular go to method is MRI image classification as they are interpretable by clinicians and can be examined after a ML model points to one claiming “that one there seems off”, this is temporary solution to deep learning techniques being powerful for classification yet uninterpretable which is a huge deficit for application in healthcare. This is why, currently, I think ML will be another tool for diagnosis not a replacement or augmentation of existing methods.

Having said that, there are other areas such as using NLP for diagnosing/detecting the onset of Alzheimer’s disease (Elif Eyigoza, 2020).

This is where I want to step in, I want to spend time to develop skills and knowledge in machine learning, leveraging my engineering experience in order to transition into ML in mental health, an area my workplace (IBM) is active in.

This project hopefully will be an introduction to that, the dataset used in this project is in a different format then what I expected, I approach it as a learning opportunity for the future.

Models Built – Model Configuration

**Data preparation**

There was very little required to prepare data for this assignment on the surface level. Originally, I was under the impression that the data consisted of MRI images, however it’s a set of correlation values and other extracted multimodal features derived from MRI scans. Datasets were combined to form one in an attempt to encompass correlation between features during feature selection.

There are no missing or incomplete records in this dataset otherwise they would need to be removed or filled in if possible. There is an option to use a scalar to reduce gaps between values however these values are already very small and close together, from running models with and without the use of scalars it has made little to no difference.

Using visualisations to see the distribution of data and a boxplot to get a five-number summary for a quick evaluation could also be used, but in the current state of this assignment there wouldn’t be much value of doing that.

The reasoning behind there being very little data preparation on the surface level is that this data set came in two files. Finding ways to better extract features would be a major To-do for future work, especially because some of these features can correlate with other illnesses

One way of working with MRI images when there is a shortage of data is to rotate them a few times to synthesis more data points for a machine learning algorithm to pick up on.

**Implementation**

**SVM/SVC**

The sigmoid kernel returns a binary value similarly to a logistic regression. The sigmoid kernel preformed best, to my limited understanding this is because it adds non-linearity when choosing which values to pass as an output and which one not to pass, if a result is > or equal to 0.51 it will round up to a 1 and likewise for 0 if its below 0.51. (Kumawat, 2019)

After careful investigation and achieving bronze via manual configuration, I was able to setup grid searching properly and got a 0.5% boost in performance with a completely different kernel, previously I used sigmoid, now its poly. The only configuration I set as default within the grid search was boosting as it caused crashes since some features only work when boosting is enabled, adding an option for false creates combinations that then cause crashes, id assume there would be a case implemented to prevent that but there isn’t.

After identifying the poly kernel as being the best, I was then able to start cutting out unneeded configurations to look at patterns, rebuild visualisations and fine tune the presplit random\_state to get better result. The entire grid\_search config has been left in place.

**Random Forest**

A similar approach was taken, a grid of params was created and evaluated against manual configurations. Feature selection was attempted but never used to retrain the model because of time constraints on my end. Visualisations are kept in the notebook.

**Linear Discriminant Analysis**

The reasoning behind using this model is that it has similar properties to SVM when using the sigmoid kernel, which showed good results, at least when compared to other models. After trying and reading the documentation on different params for each model iteration and then using grid\_search, the defaults seemed to work best with entropy as the criterion so the grid search was reverted. Currently I think more time could be spent on data modelling for these models, since parameter tunning did very little. Alternately better suited algorithms can be identified over the ones selected.

Local Evaluation – Evaluation Strategy

When comparing models and reading what each leader board score is for, I found that the private leader board score is the one that uses more data and is more relevant for the actual performance of the model. (Park, 2012).

Model evaluation consisted of:

Accuracy 🡪 fraction of predictions the model got right.

Precision 🡪 Refers is a ration of correctly predicted positives to the total of predicted positive observations.

Recall 🡪 Meanwhile recall refers to the number of those instances that are correctly predicted.

Cohen’s Kappa 🡪 This is a metric used to assess percent agreement, this is a useful metric because it also takes into account the possibility of said agreement occurring by chance. (McHugh, 2012) This is accomplished by taking the difference between the overall accuracy and the overall accuracy that can be obtained by chance. This metric is used when working with unbalanced datasets. Its an informative way to if the dataset is unbalanced or not as unbalanced datasets will produce a lower value.

Log Loss/Cross-Entropy 🡪 measures the performance of a classification model, however it requires probability values rather than binary values, for the current setup this metric was removed and not used buts its worth mentioning as it seems popular and useful for models where probability values are used.

Learning Curve 🡪 This is a visualisation that maps the change in performance as over time.

ROC 🡪 This is a graph that shows the performance of a classification model at all thresholds

AUC 🡪 This is a graph that visualises the area under the ROC curve, this can be interpreted as the probability that the model will rank a random positive example more highly than a negative example. It measures how well a prediction is ranked. (TAVISH SRIVASTAVA, 2019)

Usually for binary classification, precision, recall and F1 score should suffice. More metrics were used to get a better understanding of their application in classification models and they provide a better overview of the model itself.

K-fold Cross-Validation 🡪 data is partitioned into k number of subsamples; this is repeated k number of times and each time one of the subsets is used as the set/validation set meanwhile the k-1 subset is used to train the model.

Local Evaluation – Evaluation Results

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | Cohens Kappa | Kaggle private score | Learn-R pos or neg | AUC |
| SVM | 0.808 | 0.9 | 0.692 | 0.805 | 0.615 | 86.15% | 38- | 0.81 |
| RF | 0.808 | 0.8 | 0.727 | 0.8 | 0.601 | 74.6% | 0 | 0.71 |
| LDA | 0.769 | 0.833 | 0.714 | 0.769 | 0.541 | 82.8% | 80+ | 0.77 |

**SVM – Support Vector Machine** | **RF – Random Forest** | **LDA – Linear Discriminant Analysis**

The best predictive metric of a model’s performance from the table above is AUC derived from a ROC graph in the notebook. This is because this is what the Kaggle competition used to evaluate models. Other evaluation metrics for RF don’t seem to reflect the on Kaggle score however RF is not preferable when working with only two modalities, I hoped that because of the nature of the dataset it would perform better but it didn’t, having said that, feature selection was used but never properly implemented in a new model. SVM model excels at comparing two modalities.

Stepping to the LDA model, I think it was a step in the right direction, I wanted to explore other options for working with such a dataset and find alternatives that could outperform the SVM model. Unfortunately, I’ve ran out of time to do more work on this.

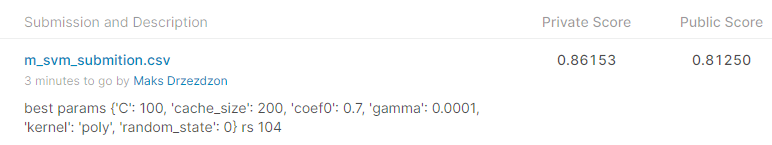
The **pos**itive or **neg**ative values for the **Learn-R**ate column stand for whether the model started performing better or worse with more data. RF had a constant value of 1 meaning there was no change with more data, meanwhile SVM showed a sharp drop at 38 using feature selection from this result would be strongly advised and explored later on, with LDA a strong progressive upward trend, potentially synthesising data could help improve the model. Both SVM and RF had the same accuracy but widely varied on every other metric. SVM had the highest score for predicting true positives at 0.9, however LDA and RF had a much higher recall meaning they missed less predictions. The F1 score takes a weighted average of precision and recall. Despite the short comings of RF and LDA, with more time and data modelling these scores could be bumped up.

After reviewing the grid\_search approach a small bump in performance was attained with a new kernel. SVM is very popular in classifying mental health disorders, rotating back again to data modelling and the use of feature selection, I think both could significantly boost the performance of SVM well in to high 80s on Kaggle’s performance report.

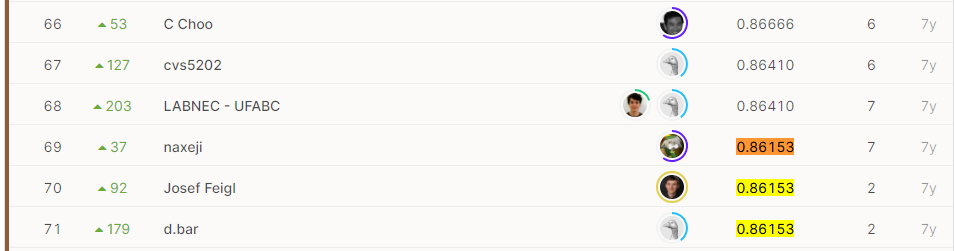
The reason for more time being required despite having done feature selection is that id have to further model the output so that it matches the training set so that Kaggle accepts the submission, otherwise I get a score of 0.

Kaggle – Performance Report

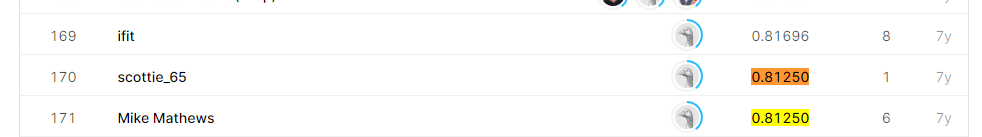
Best attempt with **support vector machine**:

**Public**: 170/313 **Private**: 69/313 - bronze

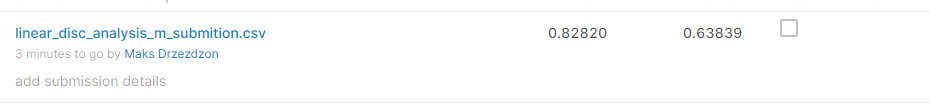
Private



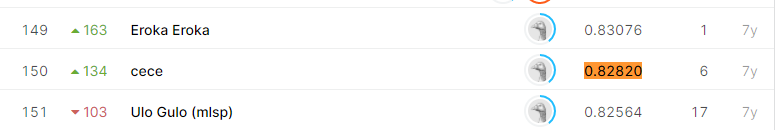
Public



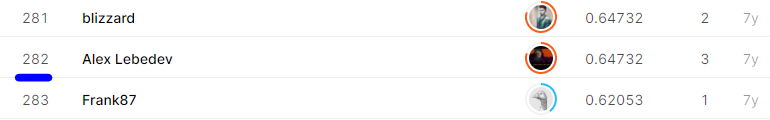
Best attempt with **discriminant linear analysis**:

**Public**: 283/313 **Private**: 150/313 

Private

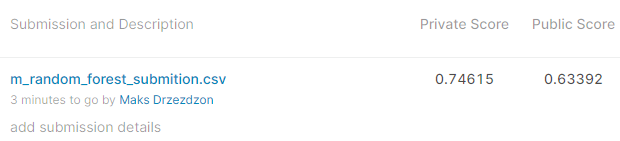


Public

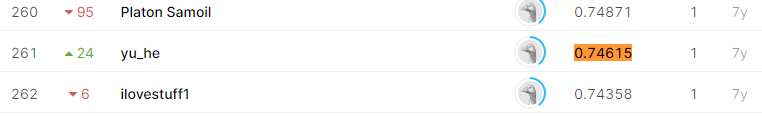


Best attempt with **random forest:**

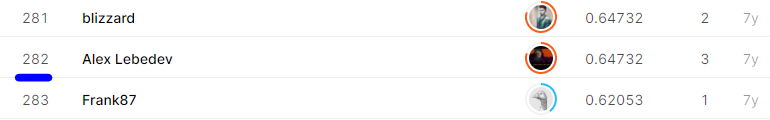
**Public**: 283/313 **Private**: 261/313



Private



Public



To-do’s – Further work

More time invested in understanding other methods of assessing model performances, similar to Cohens Kappa. Investigating other methods for handling multimodal data, along with allocating more time for feature selection. There is a lot of space for improvement which is something I’d how to achieve in my thesis using this dataset. Some methods to research later would be distance weighted discrimination function, other ways of using SVM, this can be accomplished by reading more research papers. Trying multi-layer neural networks using the previously used sigmoid function as an activation method. Id also like to reach out to the research facility and see what has come of that competition, have the models made by the winners been used, has there been any progress made since the?

# References

CONNOLLY, M. (2020, Jan 17). *adhdireland*. Retrieved from Easy-to-Miss ADHD Symptoms in Girls: https://adhdireland.ie/what-adhd-looks-like-in-girls/

David Blihar, E. D. (2020). A systematic review of the neuroanatomy of dissociative identity disorder. *Elsevier*, https://doi.org/10.1016/j.ejtd.2020.100148.

Elif Eyigoza, S. M. (2020). Linguistic markers predict onset of Alzheimer's disease. *Elsevier*, 10.1016/j.eclinm.2020.100583.

Kumawat, D. (2019, August 22). *analyticssteps*. Retrieved from 7 Types of Activation Functions in Neural Network: https://www.analyticssteps.com/blogs/7-types-activation-functions-neural-network

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276–282.

Network, M. R. (2014). *MLSP 2014 Schizophrenia Classification Challenge - Diagnose schizophrenia using multimodal features from MRI scans*. Retrieved from Kaggle: https://www.kaggle.com/c/mlsp-2014-mri

Park, G. (2012, July 6). *gregpark.io/blog/Kaggle-Psychopathy-Postmortem*. Retrieved from The dangers of overfitting: a Kaggle postmortem: http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/#:~:text=The%20public%20leaderboard%20scores%20are,subset%20of%20the%20test%20data.&text=Top%20of%20the%20private%20leaderboard,after%20the%20contest%20is%20closed.

TAVISH SRIVASTAVA. (2019, August 6). *11 Important Model Evaluation Metrics for Machine Learning Everyone should know*. Retrieved from analyticsvidhya.com/: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/